Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Florian Bäurle

Information Retrieval
WS 2012/2013

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 11

Submit until Wednesday, January 30 at 4:00pm

This exercise sheet is about comparing Naive Bayes and Support Vector Machines with respect to their ability to linearly separate the training data, and correctly classify new data.

Exercise 1 (5 points)

Classify the two-class dataset linked on the Wiki (biologists and physcisist) using your *NaiveBayesMain* program from the last exercise sheet. Again use every tenth document for training, and the rest for testing / prediction (there are no multi-label documents this time).

Exercise 2 (5 points)

In the lecture, it was explained how Naive Bayes can be seen as a linear classifier. Figure out whether, in the training phase of Exercise 1, Naive Bayes can completely separate the training data. If not, how many documents are on the wrong side? When you remove those documents from the training set (and your training data becomes linearly separable using Naive Bayes) how wide is the "band" separating the two classes.

Exercise 3 (5 points)

Extend your class *NaiveBayes* from the previous exercise sheet by a method *writeSvmLightFiles* that (after reading the file from the Wiki with *readFromCsvFile*) writes the training and test files in the format required by the *SVM Light* code from the next exercise.

Exercise 4 (5 points)

Repeat the classification task from Exercise 1, with the same division into training and test data, but this time using the *SVM Light* code from `http://svmlight.joachims.org`, as demonstrated in the lecture. Run the classification task twice: disallowing outliers (that is, going for complete linear separation) and allowing outliers. Figure out (from the SVM Light output) the band size in both cases.

Report your results in the table linked on the Wiki and briefly discuss them in your *experiences.txt* along with the usual feedback. Commit everything to our SVN, in a new subdirectory *exercise-sheet-11*.